DOCUMENT RESUME

ED 091 211 SE 017 781

AUTHOR Seymour, Lowell A.; And Others
TITLE The Measurement of Program Implementation and Students' Cognitive, Affective, and Social Performance in a Field Test of the Inquiry Role Approach (1972-73). I. Implementation: Its Documentation and Relationship to Student Inquiry Development.
PUB DATE Apr 74
NOTE 20p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (47th, Chicago, Illinois, April 1974). For related document, see SE 017 782 and 783

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS *Biology; *Educational Research; Program Evaluation; Science Education; *Secondary School Science; *Student Behavior; *Teacher Behavior; Teaching Skills
IDENTIFIERS *Inquiry Role Approach; Research Reports

ABSTRACT

This report is one of three concerning the 1972-73 field test of the Inquiry Role Approach (IRA) to biology teaching developed by the staff of the Mid-Continent Regional Educational Laboratory (McREL), Kansas City, Missouri. This paper concerns program implementation and focuses on three questions: (1) Can the IRA program be implemented? (2) Is there any significant relationship between type of training and degree of implementation? and (3) Is there any significant relationship between degree of implementation and student outcomes in biology content knowledge, cognitive inquiry skills, and affective inquiry qualities? Data were collected from 15 teachers, using a Teacher's Log, and from 1,300 students, using the Views and Preferences-C instrument and Class Activities Questionnaire. In addition, the Comprehensive Final Examination, Explorations Final Examination, Explorations in Biology, and Biology Student Behavior Inventory were used to measure student outcomes. Fourteen of the 15 teachers implemented the IRA program adequately or very adequately. No significant relationship was found between the variables, type of training and degree of implementation. Findings appeared to show that at least adequate implementation was necessary to attain development of cognitive inquiry but was not necessary for development of affective qualities and biology content knowledge. (Authors/PEB)

ED 091211

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE MEASUREMENT OF PROGRAM IMPLEMENTATION AND STUDENTS' COGNITIVE, AFFECTIVE, AND SOCIAL PERFORMANCE IN A FIELD TEST OF THE INQUIRY ROLE APPROACH (1972-73)

I. IMPLEMENTATION: ITS DOCUMENTATION AND RELATIONSHIP TO STUDENT INQUIRY DEVELOPMENT

by

Lowell A. Seymour
Lawrence F. Padberg
Richard M. Bingman
Paul G. Koutnik
Kenneth A. Burton

A paper presented to the annual meeting of the National Association for Research in Science Teaching, Chicago, 1974.

McREL projects described in this report were supported in development by funds from the U. S. Office of Education and National Institute of Education, Department of Health, Education and Welfare, under contracts OEC-3-7-062876-3076 and NE-C-00-3-0068. The opinions expressed in this paper do not necessarily reflect the position or policy of the U. S. Office of Education or National Institute of Education and no official endorsement by these agencies should be inferred.

SE017 781

In educational research and development it often happens that experimental programs or methodologies do not include within the program itself means for assuring their proper implementation. Yet if one cannot identify whether adequate implementation has occurred, there may be no real basis for deriving conclusions from any study of the program. The developers of the Inquiry Role Approach (IRA) program* considered it important that a means for assessing the implementation of IRA be included within the program materials. This would not only provide a way of measuring implementation during the field test of the program, but, more important, it would enable users of the program to measure their own implementation and to identify deficiencies to be corrected.

When plans were made for the 1972-73 field test** of the Inquiry Role Approach, attention was given to the matter of measuring implementation. Within this general problem area of measuring implementation, three specific objectives were identified: (1) To describe the degree of implementation by each participating teacher. (2) To determine whether there was a significant difference in the degree of implementation between groups of teachers receiving different types of IRA training. And (3) To determine whether there were significant differences in student outcomes in biology content knowledge, cognitive inquiry skills, and affective qualities of inquiry between students in classes where the program was inadequately implemented, adequately implemented, and very adequately implemented. This paper will discuss these three objectives from the field test.

OBJECTIVE 1: To describe the degree of implementation by each participating teacher.

The methods used to document the implementation of the Inquiry Role Approach program strike a balance between teachers' reports and students' perception of implementation. Part of the reason for this use of both students and teachers was the rationale provided by Steele, et al. (1971):

> "It was judged that the most accurate estimate of cognitive emphasis and positive learning environment could be obtained using sensitive and perceptive observers who would be in the class frequently and who were trained in using systematic procedures to collect the data. This procedure is too costly. The training, time, and support demands prohibit its use ... However, two sources of untrained observers exist in any classroom: the teacher and the students."

---

* The Inquiry Role Approach (IRA) is a method of teaching secondary biology which includes teacher training materials, teacher instructions for class use and student materials. While the goals of IRA include the learning of biology content--factual information, concepts and principles of biology--the goals emphasize inquiry skill development, social interaction skills, and attitude development necessary for good inquiry. The IRA method is based on the premise that biology content understanding, inquiry skills, social skills, and attitudes are interdependent and can be achieved best in a program that integrates them.

** For a more complete description of the IRA field test, refer to the third paper in this paper set (Seymour, et al., 1974).

Four sets of data were utilized to document the extent of implementation: (1) The per cent of activities completed by the teacher. (2) The per cent of students reaching criteria on activities completed. (3) Students' views, as measured by the Views and Preferences-C (Seymour and Bingman, 1973), of whether selected social behaviors, cognitive behaviors and class procedures characteristic of IRA were being implemented. (4) Students' views, as measured by the Class Activities Questionnaire (Steele, et al., 1971) whether selected cognitive behaviors, class procedures and teacher/student attitudes considered characteristic of IRA were present or emphasized.

The following definitions were formulated and served as the criteria for measuring extent of implementation:

Very adequate implementation - Three of the following four criteria must be met:

1) In Theme I (Teacher's Manual), 90% of activities must be completed; Theme II, 70%; Theme III, 40%.*

2) 75% of the students must reach the objectives of each activity.

3) Students will respond in the desired way on Views and Preferences-C instrument with a mean score of 3.65 or better (views items only).**

4) Students as a group (65% or more) agree at the end of Theme II that six of the following nine categories were emphasized as measured by Class Activities Questionnaire (CAQ: application, analysis, synthesis, evaluation, discussion, independence, divergence, ideas valued over grades, and enjoyment of ideas.

Adequate Implementation - Three of the four minimum criteria below must be met. The same criteria definitions are given here as were given in 1 thru 4 above with these changes:

1) Theme I, 80%; Theme II, 60%; Theme III, 10%.

2) 55%.

3) A mean score of greater than 3.5.

4) Four of the nine (CAQ) categories emphasized.

Instruments --

Data for percent of activities performed and percent of students meeting criteria on these activities was reported by the teacher using a Teacher's Log.

---

* The activities in the IRA program are grouped into three sequential sets of activities, each referred to as a Theme. Theme I has 22 activities, Theme II has 15, and Theme III has 9.

** Since V & P-C was administered as an interim measure and as a posttest the average of these two administrations was utilized for this hypothesis.

The Teacher's Log was a standard form supplied by the developers; each teacher completed a log after each activity. Information included the percent of students meeting criteria, an evaluation of the teacher's instructions in the IRA teacher's manual, an evaluation of the IRA student materials for the activity, and any alterations made in the activity procedures.

Views and Preferences-C is an instrument in which students identify whether they perceive certain behaviors or procedures being performed in the class (views items) and whether they prefer that these behaviors or procedures be performed (preference items). Students respond to items by selecting one of five choices; Strongly agree, agree, undecided, disagree, or strongly disagree. There are three sets of items dealing with social behaviors, cognitive behaviors, and class procedures. The development of the Views and Preferences-C has been reported by Seymour and Bingman (1973).

The V&P - Form C contains 50 items which were selected from 143 items of Views and Preferences - Forms A & B. The items were mainly selected on the basis of whether or not a majority of IRA students had responded in the desired direction and the items discriminated between 700 IRA and 520 non-IRA students. The non-IRA students in this sample were enrolled in BSCS biology classes and used a standard textbook laboratory approach. The data for the two groups were analyzed by calculating a chi-square for each item. The items selected, the level of significance, and the percent who chose the desired response are recorded in the Seymour and Bingman paper. Differences between the two groups were significant for 49 items and another item was retained because IRA students met the criterion level and it was deemed to measure an important aspect of the Inquiry Role Approach program.

Test-retest reliability for the Views and Preferences was found to be 0.80 (Mid-continent Regional Educational Laboratory, 1971, p. 71). More recent studies of correlations between the three sets of views items on the instrument have shown correlation coefficients of .81 (social behaviors-cognitive behaviors), .61 (social behaviors-class procedures), and .66 (cognitive behaviors-class procedures).

In determining the degree of implementation, only data from the views items were used. Responses to these items would indicate whether or not preferred IRA behaviors or procedures were being performed in the classroom as seen from the students' point of view. A mean score for each set of items can range from 1.0 to 5.0. A mean score greater than 3.50 indicates that more than 50% of the students, on the average, have responded in the preferred direction. A mean score of 3.65 or greater indicates that 65% (or more) of the students, on the average, have responded in the preferred direction.

The Class Activities Questionnaire (Steele, et al., 1971) is a second means for obtaining student feedback on whether certain activities are performed in their classes. Students respond to items by selecting one of four choices: Strongly agree, agree, disagree, or strongly disagree. The student responses are used to indicate whether eighteen factors are emphasized in the class. These factors are in turn grouped into four dimensions of class activities and climate. The four dimensions and eighteen factors measured by the CAQ are given in Table 1.

TABLE 1: Dimensions and Factors measured by Class Activities Questionnaire

DIMENSION	FACTORS
Lower Thought Processes	.Memory .Translation .Interpretation
Higher Thought Processes	.Application .Analysis .Synthesis .Evaluation
Classroom Focus	.Discussion .Test Stress .Lecture
Classroom Climate	.Enthusiasm .Independence .Divergence .Humor .Ideas Valued over Grade .Enjoyment of Ideas .Teacher Talk .Homework

Reliability of the CAQ was measured by use of the Horst formula for estimating reliability from the within class and between class variances. Reliability estimates for the four dimensions as well as sixteen factors were obtained. (Ideas Valued over Grade and Enjoyment of Ideas were not measured). Fourteen of the 20 correlations were above 0.80 with only one falling below 0.65. A study of stability of response over time indicated test-retest reliability coefficients for each of the four dimensions as 0.67, 0.91, 0.59, and 0.89, respectively.

For purposes of assessing implementation of the IRA program, nine of the eighteen factors measured by the CAQ were utilized: Application, analysis, synthesis, evaluation, discussion, independence, divergence, ideas valued over grade, and enjoyment of ideas. (In retrospect, other factors may also have been useful, particularly to show a deemphasis on memory, test stress, lecture, and teacher talk. In fact, data from these other factors generally occurred in the desired direction, that is, students agreed that these factors were deemphasized.)

Results --

Table 2 identifies the performance of each teacher in each of the four sets of data collected, and the final description of the degree of implementation achieved:

TABLE 2: The Adequacy of Implementation of IRA Classes for Each Teacher

TEACHER NO.	PERCENT OF ACTIVITIES COMPLETED BY THEME I	II	III	PERCENT OF STUDENTS REACHING CRITERIA ON ACTIVITIES COMPLETED	VIEWS & PREFERENCES-C[B] MEAN SCORE FOR "VIEWS" ITEMS	CLASS ACTIVITIES QUESTIONNAIRE[C] NUMBER OF FACTORS EMPHASIZED	ADEQUACY OF IMPLEMENTATION
01	95.4	0.0	11.1	68.5 (16)[A]	3.58	(not available)	IA[E]
02	68.2	57.1	11.1	74.7 (10)	3.69	6 (1,2,6-9)[D]	A
03	100.0	100.0	100.0	74.0 (33)	3.80	8 (1-7, 9)	VA
04	100.0	92.8	44.4	81.2 (33)	4.08	(not available)	VA
10-14[F]	77.3	85.7	11.1	76.1 (12)	3.60	6 (1-4, 6, 7)	A
20	100.0	92.8	11.1	81.8 (19)	3.96	8 (1-3, 5-9)	VA
21	100.0	92.8	11.1	78.5 (19)	3.66	5 (1, 2, 6-8)	A
22	100.0	92.8	11.1	83.1 (22)	3.80	9 (1-9)	VA
30	100.0	85.7	11.1	77.2 (22)	3.83	7 (1-7)	VA
31	100.0	85.7	11.1	84.2 (22)	3.65	(not available)	A
40	95.4	92.8	44.4	73.0 (28)	3.67	4 (1,2,5,7)	A

[A] The number of activities for which percent of students reaching criteria was reported is noted in parentheses.

[B] End of Theme I data only used for teachers 01, 03, 20, and 22; end of year data only used for teachers 04 and 21; average of interim and end of year data used for remaining teachers.

[C] End of year data used.

[D] Numbers in parentheses identify the factors emphasized: 1-application, 2-analysis, 3-synthesis, 4-evaluation, 5-discussion, 6-independence, 7-divergence, 8-ideas valued over grades, and 9-enjoyment of ideas.

[E] IA=inadequate implementation; A=adequate implementation; VA=very adequate implementation.

[F] Five teachers acting in team teaching capacity.

As Table 2 shows, five teachers achieved very adequate implementation, five achieved adequate implementation, and only one performed inadequate implementation. On-site visits and other communications with teachers tended to confirm the data--that is, professional judgement of the developers based on these communications would predict at least adequate implementation at all sites except for teacher 01. The criteria established appear to be valid, with the exception of the teachers report of the percent of students meeting criteria on each activity. This parameter for judging implementation is minimally useful unless more complete teacher reporting is obtained.

Conclusion --

Objective one was met. In our judgment it is considered both useful and necessary for meaningful evaluation of any program that the degree of implementation of the program within the experimental group be established. Further, it appears valid to utilize both student and teacher feedback as an inexpensive yet reliable means of estimating implementation.

OBJECTIVE 2: To determine whether there was a significant difference in the degree of implementation between groups of teachers receiving different types of IRA training.

HYPOTHESIS 2: There will be no significant differences between mean implementation ranks (ranking based on the four variables used to describe adequacy of implementation) for teachers receiving different types of IRA training.

The fifteen teachers participating in the IRA field test received different types of training for teaching the IRA program. The four types of training are described below:

Group 1: Teachers trained by IRA development staff, and who in turn trained the teachers in group 3.

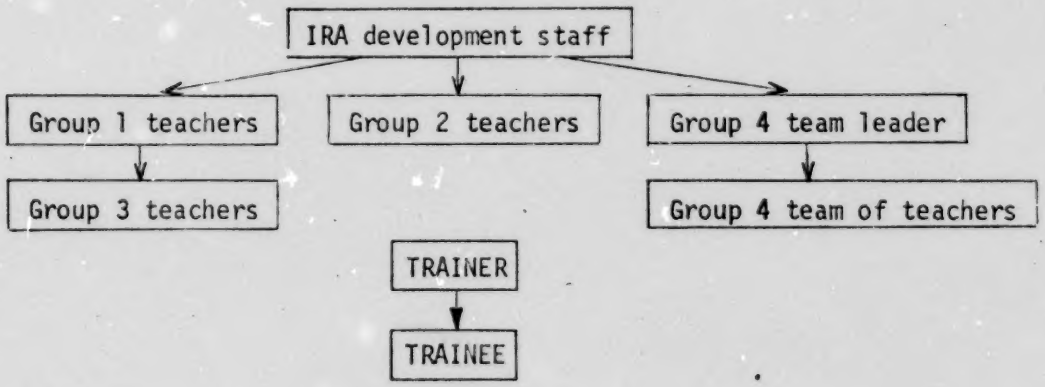
Group 2: Teachers trained by IRA development staff, and who worked independent of other IRA teachers.

Group 3: Teachers trained by Group 1 teachers.

Group 4: A group of five teachers who used team teaching. One teacher received training from IRA development staff and in turn trained the other four teachers in the group.

Figure 1 may clarify these types of training further.

FIGURE 1: Teacher training strategy for IRA field test.



In order to test hypothesis 2, a method of assigning an implementation rank was developed. The ten teachers and one teacher group were ranked on each of the four variables used to describe adequacy of implementation (percent activities completed; percent students reaching criteria; V&P-C mean score; number of CAQ categories emphasized). Each teacher's mean rank was calculated, and a final implementation rank was assigned based on the mean rank. This implementation rank is given for each teacher (or teacher group) in Table 3. A rank of 1 = lowest implementation; a rank of 11 = highest implementation.

TABLE 3: Implementation Ranks

TRAINING GROUP	TEACHER NUMBER	IMPLEMENTATION RANK	MEAN IMPLEMENTATION RANK FOR TRAINING GROUP
1	20	9.5	6.83
	30	7	
	40	4	
2	01	1	5.75
	02	3	
	03	8	
	04	11	
3	21	5	6.83
	22	9.5	
	31	6	
4	10-14	2	2.00

Mean implementation ranks were calculated for each group of teachers (grouping based on type of training); these ranks are also included on Table 3. The Kruskal-Wallace (Kruskal and Wallace, 1952) formula for determining the significance of ranked differences was applied to the mean implementation ranks for the training groups. The formula is:

$$H = \frac{12}{N(N+1)} \sum \frac{R^2}{n} - 3(N+1)$$

Results--

Teachers in Group 1, who were trained by IRA developers and who also trained other teachers in their districts, had a mean implementation rank of 6.83. Group 2, who were trained by IRA developers but did not train other teachers, had a mean implementation rank of 5.75. Group 3 teachers, who were trained by Group 1 teachers, had a mean implementation rank of 6.83. The team teaching group, Group 4, had a mean rank of 2.00. Substituting this data into the formula, the value of H is 5.86. With 3 degrees of freedom, this value is not significant at the 0.05 alpha level.

Conclusion--

The lack of statistically significant ranking differences between groups of teachers receiving different training suggests that the various training strategies used do not result in different extents of IRA program implementation. It is important that teachers (Group 3) trained by an intermediate trainer (such as those teacher/trainers in Group 1) can implement the program as well as those teachers trained directly by program developers (Groups 1 and 2) since widespread training could not depend on the relatively small group of developers as trainers. Caution must be exercised, however, in drawing conclusions, since the small sample size limits generalizability.

OBJECTIVE 3: To determine whether there were significant differences in student outcomes in biology content knowledge, cognitive inquiry skills, and affective qualities of inquiry between students in classes where the program was inadequately implemented, adequately implemented, and very adequately implemented.

HYPOTHESIS 3: There is no significant difference in student outcomes--biology content knowledge, cognitive inquiry skills, and affective qualities of inquiry--for students in classes with different degrees of implementation.

Method/Procedures--

Data used to test the null hypothesis came from student posttests (students of the 15 IRA teachers discussed under Objectives 1 and 2) administered in late May and early June of 1973. These posttests measured the following student outcomes:

1) Comprehensive Final Examination-Forms J & K (CFE)(Biological Sciences Curriculum Study, 1965), used to measure biology achievement.

2) Explorations in Biology-Topic 1 (EIB-1)(Koos, et al., 1972), used to measure students' ability to formulate a hypothesis, design a study, interpret data or findings, and synthesize knowledge gained from the investigation.

3) Biology Student Behavior Inventory (BSBI)(Steiner, 1970), used to measure students' curiosity, openness, satisfaction, and responsibility.

Instruments--

Comprehensive Final Examination (CFE)

The CFE (see Biological Sciences Curriculum Study, 1966) is designed as a comprehensive examination of the achievement in biology attained by students in a first-year secondary level biology class. Specifically the instrument has been designed for the BSCS courses using any of the three BSCS textbooks; however, it is seen as applicable to other modern biology curricula as well. Two equivalent forms, J and K, have been developed.

Validity: The validity of the CFE has been primarily determined by the judgment of subject matter specialists and the supervisors of the writing teams for the three BSCS texts. In this manner the instrument has been judged to be valid in terms of covering the content of the three texts. In addition, validity was studied by determining the correlation between student scores on the CFE and on each of the four Quarterly Achievement Tests designed to accompany the three text versions. The coefficients of correlation range from .63 to .82.

Reliability: Both internal consistency of each form and correlation between forms have been studied. Using the Kuder-Richardson 20 procedure (Kuder and Richardson, 1937) with a sample of 740 cases, coefficients of internal consistency ranging from .76 to .86 were found, with a median coefficient of .82 for Form J and a median coefficient of .84 for Form K. Coefficients of correlation for scores obtained on Form J and Form K have been found to range from .72 to .85 with a median coefficient of .79 ($N = 2500$).

Biology Student Behavior Inventory (BSBI)

The BSBI (see Steiner, 1970) is a 39-item instrument designed to measure the frequency of occurrence of specific student behaviors indicative of four attitudes considered necessary for cognitive inquiry -- curiosity, openness, satisfaction and responsibility. The student is presented with a situation and a selection of possible behaviors or actions that could be taken in that situation. The student indicates what he would do in this situation by selecting one behavior. The preferred responses (which receive a score) are behaviors indicative of one of the four attitudes given above. Four subscores or subscales are therefore determined. 11 of the items are used to determine the curiosity subscores; 17 are used for the openness subscore; 7 for the satisfaction subscore; and 4 for the responsibility subscore.

Validity: Validity has been studied in three ways--by a panel of nine judges; by correlation of student item scores with student subscore scores (this was used primarily to confirm categorization when judges did not show a high percentage of agreement); and by correlations with a second instrument (Observational Record of Affective Behaviors, ORAB) which measured the same attitudes utilizing fewer behaviors and an observational approach.

The judges' agreement has been reported as the percent agreeing with the test author. In keying the BSBI items to one of the four attitudes, 67 percent agreement or higher was found for 33 of the 39 items; average percent agreement for all 39 items was 83 percent.

To confirm the judges' findings and in particular to evaluate the categorization of the six items which showed low (below 67 percent) percentages of agreement, a Pearson product-moment correlation coefficient was determined for each item. Student item scores were correlated with each of the four student subscale scores. This process confirmed the validity of the previous categorization of items.

Finally Pearson coefficients were found for student scores on three subscales of the BSBI (curiosity, openness and responsibility) and total BSBI scores (using only three subscales) correlated with the same three subscores on the ORAB and the total ORAB score. The curiosity subscales had a correlation coefficient of -.45; the openness subscales, .88; the responsibility subscales, .75; the total scores, .83 (for significance at the .05 level, $r \geq .75$). The low curiosity subscale correlation appeared to be due to the fact that the ORAB measured primarily only one behavior indicating curiosity while the BSBI measured five behaviors; thus the two instruments were not measuring the same behaviors and low correlation could be expected. It should also be noted that the ORAB contained a non-inquiry subscale; this subscale showed negative correlation with each BSBI subscale and the BSBI total score.

Reliability: An estimate of the reliability of each subscale was determined using a split-half technique. Pearson product-moment coefficients of correlation were found and adjusted using the Spearman-Brown formula (Guilford, 1965, p. 457).

With a student N of 1153, the following values were computed: curiosity, $r = .67$; openness, $r = .68$; satisfaction, $r = .71$; responsibility, $r = .37$; for significance at the .01 level, $r \geq .07$. With a class N of 48, the following values were computed: curiosity, $r = .78$; openness, $r = .68$; satisfaction, $r = .86$; responsibility, $r = .51$; for significance at the .01 level, $r \geq .37$. A Cronbach alpha (Cronbach, 1951) was also computed to determine internal consistency for each subscale.

The alpha values: curiosity, α = .65; openness, α = .71; satisfaction, α = .66; responsibility, α = .43 (N = 1153). BSBI A, B, C, D, and total had Cronbach alpha values of 0.55, 0.78, 0.68, 0.37, and 0.84 respectively. The sample was the experimental group.

Explorations in Biology (EIB)

The EIB series (see Koos, 1970, 1971, 1972, and Koos and Chan, 1972) is a set of eight simulated problem-solving instruments designed to measure cognitive inquiry skills. These instruments have been developed in the period of 1969-72 as a component of the Development of Inquiry Skills Program of McREL. The instruments are designed to measure the following inquiry skills:

14 Inquiry Objectives -

1. Identifying a phenomenon to investigate.
2. Identifying the question arising from the identification of this phenomenon.
3a. From a list of readings, selecting and evaluating reports possibly yielding useful information about the event noted. (Explorations 2, 3, 4, 5, and 6)*
3b. From relevant readings on the problem presented, decide if given hypotheses are tenable. (Explorations 1, 7, and 8)*
4. Differentiating likely causes of this event from unlikely causes.
5. Selecting a single hypothesis to investigate.
6. Selecting an array of methods appropriate to the investigation.
7. Identifying the independent variable to be studied.
8. Identifying conditions required for conducting a laboratory study on this topic.
9. Choosing a plan which would yield data affording a test of the hypothesis.
10. Identifying assumptions necessary for interpretation of data resulting from carrying out the plan.
11. Identifying the data which would result from carrying out this plan.
12. Identifying justifiable conclusions from data associated with a class experiment on this topic.

* EIB's 1 through 6 were developed to measure the above set of objectives including 3a but not 3b. The format for EIB's 7 and 8 was slightly changed from an earlier format used for EIB's 1 thru 6. With this new format, Objective 3b was substituted for 3a. In 1972 EIB 1 was revised into the new format.

13. From a heterogeneous list of questions, identifying new questions which might arise as a result of carrying out this investigation.

14. Integrating results of this study with those reported by other investigators in related areas.

In the 1972-73 field test, EIB 1 was used as the pre and posttesting instrument for assessing cognitive inquiry skill student outcomes and pre-to-post gain.

Validity: Objectives were selected for the EIB's based on studies by Burmester (1952), Kaplan (1967), and Suchman (1962). With the completion of the detailed McREL-BSCS set of inquiry objectives (Bingman, et al., 1969), studies were made to learn the extent to which EIB items would be referenced to similar objectives listed in the Inquiry Objectives in the Teaching of Biology document. These studies were previously reported by Koos (1970).

Changes in the Explorations in Biology since these studies were undertaken have been primarily format changes and changes in wording to clarify directions and meaning. However, the inquiry objective content validity was reviewed in the summer of 1972 by two McREL staff members and a teacher-consultant.

Working independently each judge keyed the test items using: the 14 EIB objectives, a category for items in which sequence of test steps were chosen, and a category for items not related to any of the objectives or step choice. Disagreements were found for less than 15 percent of the items. In most cases, disagreements resulted from misreading, or misinterpreting, the test items or directions. In all cases, disagreements were discussed and concensus reached for keying the item. In addition, the EIB objectives were categorized by the judges as related to six major areas of cognitive inquiry behaviors. Table 4 presents the categorization of the objectives and the items in EIB 1 keyed to each objective.

TABLE 4: EIB 1 Items Keyed to Inquiry Objectives

		ITEMS ON EIB 1-A	ITEMS ON EIB 1-B
AREA I - Formulating a Problem	Objective 1	1	
	Objective 2	50	
AREA II - Searching for Information	Objective 3a		
AREA III - Formulating Hypotheses	Objective 3b	8-17	
	Objective 4	51-60	
	Objective 5	48,49	
AREA IV - Designing an Experimental Study	Objective 6	36-45	
	Objective 7		1
	Objective 8		2-6
	Objective 9	46,47	17,43-47
AREA V - Interpreting the Data or Findings	Objective 10		7-16
	Objective 11		18-42*
			48-57
	Objective 12		58-67
AREA VI - Applying and Synthesizing Knowledge	Objective 13		68-77
	Objective 14		78
Step choice items	(not scored)	2-7	

* Students choose one set of 5 items to respond to in the 18-42 group

Based on this categorization and assignment of test items to objectives, subscores for each of the six inquiry areas can also be determined in scoring the EIB's.

Construct validation studies have been made to compare EIB 1 with BSCS Comprehensive Final Examination, Differential Aptitude Test-Verbal Reasoning & Numerical Ability and Abstract Reasoning (Bennett, et al., 1959), California Basic Skills Test (Tiegs and Clark, 1955-56), Iowa Tests of Basic Skills, (Lindquist and Hieronymus, 1955-56), Scholastic High School Placement Test (Anderhalter, et al., 1959), and Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1961). A Pearson product-moment correlation of .53 was found between EIB 1 and DAT-Abstract Reasoning and Watson-Glaser Critical Thinking. Other correlations were found to be very low. "...construct validity is offered for those EIB 1 items which tap cognitive operations involving verbal formulation of biological problems, verbal interpretation of non-verbal data, and analysis of quantitative information presented in tabular or graphic form. This suggests that the intellectual factors of verbal reasoning and numerical ability are factors basic to successful inquiry" (Koos, 1970, p. 15).

Reliability: The developmental 1969 version of EIB 1 was shown to have a coefficient of internal consistency (Kuder Richardson 20 procedure - Kuder and Richardson, 1937), of .96 when tested with a heterogeneous group of 451 students; .74 when tested with a more homogeneous group of 150 students.

The later 1970 versions of EIB 1 and 2 were tested on several occasions in the spring of 1970 and in the 1970-71 school year. Coefficients of internal consistency (Cronbach alpha, Cronbach, 1951) ranging from .40 to .86 and averaging from .75 to .99 and averaging .87 were found for EIB 2.

While reliability was adequately demonstrated by these analyses, the EIB 1 format was revised during the summer of 1972. Major changes involved the items keyed to Objectives 3a and 3b. Objective 3b was substituted for 3a, and related items were revised or replaced. This change was made to insure that all students were provided the same background information on the topic; formerly, readings from related science literature were optional. In addition, the items keyed to Objective 4 were reduced from 20 to 10. Some item numbering changes in Part A were also made. In order to establish the degree of reliability of the 1972 revised instrument, coefficients of internal consistency (Cronbach alpha) were determined for the total scores and part scores I, III, IV, V and VI of EIB 1 using the posttesting data from the IRA field test students. The coefficients are presented in the following table.

TABLE 5: EIB Coefficient of Internal Consistency

	COEFFICIENT OF INTERNAL CONSISTENCY	N
EIB-1, Part I	0.24	1,005
EIB-1, Part III	0.62	1,005
EIB-1, Part IV	0.83	1,005
EIB-1, Part V	0.85	1,005
EIB-1, Part VI	0.88	1,005
EIB-1, Total score	0.87	1,005

It should be noted that the scoring key for EIB 1 was revised during 1972. Previous scoring keys had been devised by the principal test author and had not been reviewed by others. In discussing aspects of the 1972 revision of EIB 1, it was found that IRA staff members disagreed with the suggested scoring of some items. A more thorough review was planned with five McREL staff members acting as judges. New scoring keys, reflecting consensus among the five judges, were developed. The degree of difference between the original author's key and the revised key can be determined from the following table.

TABLE 6: Comparison of EIB-1 Scoring on Original and Revised Scoring Keys

	(1) TOTAL POSSIBLE RESPONSES	(2) SCORED RESPONSES, ORIGINAL KEY	(3) SCORED RESPONSES UNCHANGED	(4) SCORED RESPONSES, SCORING DELETED	(5) UNSCORED RESPONSES UNCHANGED	(6) UNSCORED RESPONSES, SCORING ADDED	(7) PERCENT AGREEMENT $\% = \frac{(3)+(5)}{(1)} \times 100$
EIB-1A	158	51	34	17	94	13	$\frac{34+94}{158} \times 100 = 81.0\%$
EIB-1B*	220	82	76	6	128	10	$\frac{76+128}{220} \times 100 = 92.7\%$
TOTAL	378	133	110	23	222	23	$\frac{110+222}{378} \times 100 = 87.8\%$

* All optional items included.

The author's key has used a "weighted" scoring system. Scored responses could be awarded either 2 or 1 point. Criteria for weighting the value of responses appeared to include difficulty of the item, degree of accuracy of response (when more than one response to an item was scored), and whether the response was negative rather than positive (the author felt a negative response to an item was psychologically more difficult to make).

The panel of judges felt that these criteria were not consistently applied. The author has not specified a systematic approach for assigning weighted scores. The judges, therefore, decided to delete weighting of scores as much as possible--weighted scores are used in the revised key only for optional sections when necessary to maintain equal chance scores for each option presented.

All EIB 1 data in this report utilizes the revised key. Maximum and chance scores for EIB 1 total score and part scores are given in the following table.

TABLE 7: EIB Scoring Key Changes

	ORIGINAL AUTHOR KEY EIB-1		REVISED PANEL KEY EIB-1	
	MAXIMUM SCORE	CHANCE SCORE	MAXIMUM SCORE	CHANCE SCORE
Part I	4	1.20	2	.40
Part III	41	14.08	22	7.07
Part IV	45	17.15	24	11.30
Part V	60	27.00	40	17.50
Part VI	21	10.10	10	4.70
TOTAL	171	69.53	98	40.37

Data Analysis/Results: Eleven student outcome variables were identified: CFE total score, EIB-I total score and four subscale scores, and BSBI total score and four subscale scores. An analysis of covariance was computed for each of the eleven student outcome variables (note that EIB-subscale I was not used due to subscale unreliability as discussed previously). Pretest scores were held constant for each variable analyzed. The Newman-Keuls statistical test was used to determine which pairwise differences were significant.

Table 8 presents the adjusted posttest means and F ratios for comparing student outcome variables for the three subgroups based on degree of implementation. Table 9 presents the results of the Newman-Keuls analysis.

TABLE 8: Adjusted Means and F Ratios for Comparing Subgroups Based on Degree of Implementation

VARIABLE	INADEQUATE IMPLEMENTATION		ADEQUATE IMPLEMENTATION		VERY ADEQUATE IMPLEMENTATION		F RATIO	DF
	ADJUSTED MEAN	N	ADJUSTED MEAN	N	ADJUSTED MEAN	N		
EIB III	9.48	22	11.69	204	12.11	129	5.81*	(2,351)
EIB IV	14.75	23	17.95	144	18.24	114	11.62*	(2,277)
EIB V	18.52	25	23.89	202	25.33	129	13.46*	(2,353)
EIB VI	6.35	25	6.96	180	7.57	117	5.57*	(2,330)
EIB Total	49.14	23	62.35	149	64.38	114	21.99*	(2,277)
BSBI A	2.79	58	2.65	144	2.76	114	1.84	(2,312)
BSBI B	3.67	58	3.64	144	3.70	114	.30	(2,312)
BSBI C	3.54	58	3.51	144	3.61	114	.92	(2,312)
BSBI D	3.65	58	3.76	144	3.78	114	.481	(2,312)
BSBI Total	13.71	58	13.56	144	13.80	114	.493	(2,312)
CFE	20.20	59	19.17	197	20.38	93	1.68	(2,345)

* Sig. at the .01 level.

TABLE 9: Newman-Keuls Post Hoc Analysis for Extent of Implementation

	IA / A	IA / VA	A / VA
EIB III	*	*	
EIB IV	*	*	
EIB V	*	*	
EIB VI		*	
EIB Total	*	*	

* Significant at .01 level

Five of the eleven F ratios are significant at the .01 level of significance. These involved the following student outcome variables: EIB III, EIB IV, EIB V, EIB VI, and EIB total score.

For the EIB III comparisons, the Newman-Keuls post hoc analysis indicated that the achievement level of the students under the teacher with inadequate implementation was significantly below both the other subgroups. For the EIB IV scores, the post hoc test indicated that the students under the inadequate implementation teacher were significantly lower than both the other subgroups. The same pattern is true for the EIB V and EIB total score. For EIB VI only the very adequate and inadequate means were significantly different. All of the comparisons were significant at the .01 level of significance.

Interpretation: The data presented suggests that at least adequate implementation is necessary to attain development of cognitive inquiry, but not necessary for development of affective qualities and biology content knowledge. Much caution must be exercised in interpreting this data. Data from only one teacher is included in the "inadequate implementation" category. Further, the students in this teacher's classes were all ninth grade students (compared to primarily tenth grade students in adequately and very adequately implemented classes); and students were in class only 180 minutes/week. This teacher strongly emphasized social and attitudinal development (note that there was no significant difference between this teacher's class and all other classes in the area of affective qualities). This emphasis contributed to the lack of use of much of the IRA program materials (no activities in Theme II and only 11 percent of Theme III activities were completed). The lack of completion of IRA activities may have strongly contributed to the significantly lower cognitive inquiry scores. Further studies using more carefully controlled groups (in terms of grade level, class structure, etc.) and larger sample size might give more conclusive results.

The question is also raised as to the relative validity of the four variables used to evaluate degree of implementation. It may be appropriate to place greater emphasis on certain variables (for example, percent of IRA activities completed) than on others.

In order to further clarify the possible relationships between student otucomes and degrees of implementation, correlation coefficients were computed between each of the eleven measures of student outcomes and the three degrees of implementation. Unadjusted posttest scores were used, thus allowing for larger numbers of students included in the data than in the analysis of covariance reported above. The results of this analysis are presented in Table 10.

TABLE 10: Correlations Between Type of Implementation and Student Outcome Variables

Variable			r	n
EIB	III	Formulate Hypotheses	.052	840
EIB	IV	Design a Study	.117**	703
EIB	V	Interpret Data	.206**	836
EIB	VI	Synthesize Knowledge	.120**	814
EIB		Total Score	.168**	703
BSBI	A	Curiosity	.101*	593
BSBI	B	Openness	.201**	593
BSBI	C	Satisfaction	.160**	593
BSBI	D	Responsibility	.203**	593
BSBI		Total Score	.227**	593
CFE			.129**	804

* Significant at the .05 level
** Significant at the .01 level

These low correlations indicate that a meaningful linear relationship between degrees of implementation and student outcomes is not substantiated. However, the low correlations may be attributed to the lack of significant differences in student outcome variables between adequately and very adequately implemented classes.

Summary--

During the 1972-73 field test of the Inquiry Role Approach, degree of implementation of the program was measured by (1) number of activities each teacher performed, (2) percent of students meeting objectives of those activities performed, (3) students views of classroom behaviors as measured by the Views and Preferences-C, and (4) student views of classroom behaviors as measured by the Class Activities Questionnaire. Adequate implementation of the IRA program has been documented for fourteen of the fifteen teachers using IRA in the 1972-73 field test.

Teachers were either trained directly by IRA development staff members, or by other teachers who had received training from the IRA staff. Differences in adequacy of implementation do not appear to be related to the types of IRA training.

Significant differences in student outcomes were found between the students in classes where IRA was not adequately implemented, and the students in classes where IRA was adequately implemented. Caution must be used in drawing conclusions since only one teacher inadequately implemented the program.

Some questions have been raised as to the validity or relative importance of each of the four factors used to determine implementation adequacy.

REFERENCES

Anderhalter, O. P., Gawkoski, R. A., and O'Brien, J. Scholastic high school placement test. Bensonville, Illinois: Scholastic Testing Service, 1959.

Bennett, G. K., Seashore, H. G., and Wesman, A. G. Differential aptitude tests. New York: The Psychological Corporation, 1959.

Bingman, R. M., Ed., Anderson, J. R., Blankenship, J. W., Carter, J. L., Cleaver, T. J., Jones, W. G., Kennedy, M. H., Klinckmann, E., Koutnik, P. G., Lee. A. E., and Stothart, J. R. Inquiry objectives in the teaching of biology. Kansas City, Missouri: Mid-continent Regional Educational Laboratory, 1969.

Biological Sciences Curriculum Study. Comprehensive final examination, forms J & K. New York: The Psychological Corporation, 1965.

Biological Sciences Curriculum Study. Manual for the Comprehensive Final Examination in first-year biology. New York: The Psychological Corporation, 1966.

Burmester, M. A. Behavior involved in the critical aspects of scientific thinking. Science Education, 1952, 36, 259.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297.

Guilford, J. P. Fundamental statistics in psychology and education. New York: McGraw-Hill Book Co., 1965.

Kaplan, E. H. The Burmester Test of aspects of scientific thinking as a means of teaching the mechanics of the scientific method. Science Education, 1967, 51, 353.

Koos, E. M., Burmester, M. A., Garth, R. E., and Stothart, J. R. Explorations in biology: Topic 1. Bird population. Kansas City, Missouri: Mid-continent Regional Educational Laboratory, 1972.

Koos, E. M., and Chan, J. Y. Technical report no. 3. Criterion-referenced tests in biology. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Koos, E. M. Technical report no. 1. A report on developmental studies of a series of measures of inquiry skills in biology, Explorations in Biology. Kansas City, Missouri: Mid-continent Regional Educational Laboratory, 1970.

Koos, E. M. Technical report no. 2. A report on developmental studies of a series of measures of inquiry skills in biology, Explorations in Biology. Kansas City, Missouri: Mid-continent Regional Educational Laboratory, 1971.

Koos, E. M. Technical report no. 4. A report on developmental studies of a series of measures of inquiry skills in biology, Explorations in Biology. Kansas City, Missouri: Mid-continent Regional Educational Laboratory, 1972.

Kruskal, W., and Wallis, W. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 1952, 47, 583.

Kuder, G. F., and Richardson, M. W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151.

Lindquist, E. C., and Hieronymus, A. N. Iowa tests of basic skills. New York: Houghton Mifflin Company, 1955-56.

Mid-continent Regional Educational Laboratory. Annual report to the U. S. Office of Education. Kansas City, Missouri: Mid-continent Regional Educational Laboratory, 1971.

Seymour, L. A., and Bingman, R. M. Development of Views and Preferences - C. A paper presented to the annual meeting of the National Association for Research in Science Teaching, Detroit, 1973.

Seymour, L. A., Bingman, R. M., Koutnik, P. G., Padberg, L. F., Havlicek, L. L., Kocher, A. T., and Burton, K. A. The measurement of program implementation and students' cognitive, affective, and social performance in a field test of the Inquiry Role Approach (1972-73). III. Evaluation of the Inquiry Role Approach methodology. A paper presented to the annual meeting of the National Association for Research in Science Teaching, Chicago, 1974.

Steele, J., House, E., and Kerins, T. An instrument for assessing instructional climate through low inference student judgments. American Educational Research Journal, 1971, 8, 449.

Steiner, H. E. A study of the relationship between teacher practices and student performance of selected inquiry process behaviors in the affective domain in high school biology classes. Unpublished doctoral dissertation, University of Texas, Austin, 1970.

Suchman, J. R. The Elementary School Training Program in scientific inquiry. Project 216, U. S. Office of Education. University of Illinois, Urbana, 1962.

Tiegs, E. W., and Clark, W. W. California basic skills test. New York: Houghton Mifflin Company, 1955-56.

Watson, G., and Glaser, E. M. Watson-Glaser critical thinking appraisal, Form YM. New York: Harcourt, Brace and World, 1961.